Published online 2 May 2008 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* ***W475–W480***
doi:10.1093/nar/gkn240

OGtree: a tool for creating genome trees of prokaryotes based on overlapping genes

Li-Wei Jiang[1], Kuang-Lun Lin[1] and Chin Lung Lu[1,2,*]

[1]Institute of Bioinformatics and [2]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan

Received February 2, 2008; Revised April 10, 2008; Accepted April 15, 2008

ABSTRACT

OGtree is a web-based tool for constructing genome trees of prokaryotic species based on a measure of combining overlapping-gene content and overlapping-gene order in their whole genomes. The overlapping genes (OGs) are defined as adjacent genes whose coding sequences overlap partially or entirely. In fact, OGs are ubiquitous in microbial genomes and more conserved between species than non-OGs. Based on these properties, it has been suggested that OGs can serve as better phylogenetic characters than non-OGs for reconstructing the evolutionary relationships among microbial genomes. OGtree takes the accession numbers of prokaryotic genomes as its input. It then downloads their complete genomes from the National Centre for Biotechnology Information and identifies OGs in each genome and their orthologous OGs in other genomes. Next, OGtree computes an overlapping-gene distance between each pair of input genomes based on a combination of their OG content and orthologous OG order. Finally, it utilizes distance-based methods of building tree to reconstruct the genome trees of input prokaryotic genomes according to their pairwise OG distance. OGtree is available online at http://bioalgorithm.life.nctu.edu.tw/OGtree/.

INTRODUCTION

The increasing availability of complete prokaryotic genomes provides us with an opportunity to reconstruct their genome trees based on the whole genomic information of organisms rather than based on individual genes or a small number of genes. In addition to sequence-based phylogenomic approaches, methods based on whole genomes, like those based on gene content (i.e. the presence and absence of genes) (1,2) and gene orders (3–5) can be used to construct more precise and robust phylogenetic trees that are less influenced by anomalous events. As pointed out in refs (6,7) however, the genome trees constructed only based on gene content or gene order may not be suitable for microbial genomes, because gene content (respectively, gene order) might have changed too little (respectively, too much) for biologists to perform adequate analyses of evolutionary distances between closely (respectively, distantly) related genomes. More recently, to address these problems, Luo *et al*. (6,7) have proposed an alternative way to reconstruct genome trees of bacteria based on the presence and absence of overlapping genes (OGs) in their complete genomes.

The OGs are defined as adjacent genes whose coding sequences partially or entirely overlap. OGs are ubiquitous in microbial genomes, because approximately a third of all genes in all the microbial genomes sequenced to date are overlapping (8,9). In fact, there is a strong relationship between the total number of genes and the number of OGs (8,9). In addition, it has been reported that OGs are more conserved between species than non-overlapping genes (10–12), because a mutation in the overlapping region causes changes in both genes and therefore natural selection against such mutations should be stronger. Based on these properties, Luo *et al*. (6,7) have reported that OGs can serve as better phylogenetic characters than non-OGs for reconstructing the evolutionary relationships among bacterial genomes.

For the phylogenetic reconstruction of bacterial genomes, Luo *et al*. (6) defined the *orthologous OG pairs* between two different genomes, say *i* and *j*, to be pairs of genes that overlap in genome *i* and have orthologous counterparts that overlap in genome *j*. In an analogous method to that used in the analysis of gene content, they defined a new distance measure between two genomes based on the normalized number of their shared orthologous OG pairs. Based on this definition, they utilized current distance-based approaches of building tree, such as neighbor-joining (NJ) and unweighted pair-group method using arithmetic averages (UPGMA), to construct the genome trees of many completely sequenced bacterial genomes. In addition, Luo *et al*. (7) have further maintained an interactive database server, called

[*]To whom correspondence should be addressed. Tel: +886 3 5712121 ext. 56949; Fax: +886 3 5729288; Email: cllu@mail.nctu.edu.tw

© 2008 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

BPhyOG (http://cmb.bnu.edu.cn/BPhyOG/), which allows the user to browse the genome trees of some bacterial genomes that were calculated in advance on the basis of shared orthologous OG pairs. However, their genome trees are not greatly consistent with those produced by traditional phylogenetic approaches based on 16S rRNAs and concatenation of multiple proteins (refer to the Experiments section for details).

In fact, during evolutionary process, species genomes are subject to genome rearrangements (e.g. reversals and transpositions) that alter the order and orientation of genes on the genomes, leading to that the orders of orthologous genes, as well as the ones of orthologous OG pairs certainly, even between two closely related species may not be conserved. This suggests that not only OG content but also orthologous OG order should be considered to reconstruct the genome trees of prokaryotic species. For this purpose, we define the OG distance between two genomes based on a measure of combining OG content and order in their whole genomes (refer to the Methods section for its detailed definition). We then use UPGMA, as well as NJ and FM (Fitch–Margolias), to build the genome tree of prokaryotic genomes according to their pairwise OG distance.

We have developed a web-based tool, called OGtree (http://bioalgorithm.life.nctu.edu.tw/OGtree/), for constructing the genome trees of prokaryotes based on OG distance between prokaryotic complete genomes. In addition, we have tested our OGtree on several Proteobacteria complete genomes to assess its quality of genome tree reconstruction. Compared with the phylogenetic trees produced by Luo *et al.* (6,7) the genome trees constructed by our OGtree are quite consistent with those reference trees that were reconstructed based on 16S rRNAs as well as concatenation of multiple proteins. All these results have suggested that our OGtree can serve as a useful tool for constructing more precise and robust genome trees for prokaryotic genomes.

METHODS

As used in the studies of genome rearrangements, we utilize a signed integer to represent a gene encoded in a chromosome, with its sign indicating the transcriptional orientation of the corresponding gene (e.g. '+' stands for $5' \rightarrow 3'$ and '−' stands for $3' \leftarrow 5'$). Moreover, we use a pair of signed integers (x, y) to represent an OG of x and y. Basically, there are three possible overlapping types (or structures/directions) of OGs (11,13): (i) *unidirectional* OGs with sign (+,+) or (−,−), that is, the 3′ end of one gene overlaps with the 5′ end of the other, (ii) *convergent* OGs with sign (+,−), that is, the 3′ ends of the two genes overlap and (iii) *divergent* OGs with sign (−,+), that is, the 5′ ends of the two genes overlap. It has been reported that in prokaryotic genomes unidirectional OGs are most widespread, convergent OGs are less common and divergent OGs are rare (8,9,13).

For our purpose, the orthologous OG pairs we considered here are further restricted to those orthologous OG pairs with the same (i.e. conserved) overlapping structures. Let $\{c_1, c_2, \ldots, c_n\}$ denote the set of total orthologous OG pairs between two given genomes G_i and G_j. Then we represent these two genomes by two permutations $G_i = (a_1, a_2, \ldots, a_n)$ and $G_j = (b_1, b_2, \ldots, b_n)$, respectively, on the same set of $\{c_1, c_2, \ldots, c_n\}$. We also say that, for example, a_k precedes a_{k+1} in genome G_i, where $1 \leq k < n$, and a_n precedes a_1 if G_i is circular. For simplicity of our description, we here assume G_i and G_j to be circular, because the genomes of prokaryotes are typically circular. Two consecutive OGs, say (u, v) and (x, y) with (u, v) preceding (x, y), in G_i determine a *breakpoint* if neither (u, v) precedes (x, y) nor $(-y, -x)$ precedes $(-v, -u)$ in G_j. It is not hard to see that the number of breakpoints in G_i is equal to the number of breakpoints in G_j. Then, we define the *overlapping-gene distance* $D_{i,j}$ between G_i and G_j as follows.

$$D_{i,j} = w_o \times \left(\frac{b_{i,j}}{n}\right) + w_c \times \left(\frac{x_i - n}{x_i} + \frac{x_j - n}{x_j}\right)$$

In the above formula, $b_{i,j}$ denotes the number of breakpoints in genome G_i with respect to genome G_j, and x_i and x_j denote the numbers of total OGs in G_i and G_j, respectively. Note that if the considered genomes are linear, the denominator of the first term in the right hand of this equation should be $n - 1$, because in this case it is the maximum number of breakpoints between G_i and G_j. Basically, $D_{i,j}$ evaluates the distance between G_i and G_j by considering the orthologous OG order measure as defined in the first term (i.e. the normalized breakpoint distance) and the OG content measure as defined in the second term (i.e. the sum of the ratios of OGs found in one genome but not found in another genome to the number of total OGs found in a genome). Then w_o and w_c can be considered as the weight of orthologous OG order and the weight of OG content, respectively, where both of their defaults are 1's in OGtree.

Subsequently, we describe the details about the procedures we used to develop our OGtree. The first step is to download complete genomes from the National Centre for Biotechnology Information (NCBI) according to the accession numbers specified by the user. The putative genes are then extracted from each of these genomes on the basis of the coding sequence (CDS) annotation. Inevitably, some of these putative genes may be misannotated in each genome downloaded from the NCBI. We may therefore exclude those genes that were annotated as being unknown, hypothetical or putative for a stringent analysis. In addition, horizontal gene transfer (HGT), the transfer of genes between different species, has been reported to be very common in prokaryotes (14). It may obscure the OG pairs with which we hope to reconstruct the genome tree of prokaryotes. Hence, we offer an additional option in our OGtree to remove those genes that were annotated as horizontally transferred genes at the HGT-DB database (14), where HGT-DB currently provides the lists of putative horizontally transferred genes for a large number of prokaryotic complete genomes.

Next, we use BLASTP program to determine putative orthologous genes between two genomes by using bidirectional best hit (BBH) approach. A BBH is defined to be a pair of genes *a* and *b* from two genomes G_i and G_j such that *b* is the best hit (i.e. most similar gene), when *a* is compared against all genes of G_j and vice versa. It has been evidenced that such a BBH approach of identifying putative orthologs works reasonably well for bacterial genomes (15). In addition, we use Inparanoid (14) as an alternative to identify putative orthologous genes between any two genomes. It has been demonstrated that Inparanoid is the best among five currently existing methods of automatically detecting orthologous genes (16). After that, two adjacent genes in each genome are identified as OGs, or an OG pair, if their CDSs overlap partially or completely. Two OGs, say (a, c) and (b, d), from different genomes are then considered as an orthologous OG pair if *a* and *b*, as well as *c* and *d*, are orthologous to each other, and (a, c) and (b, d) have the same overlapping structure.

Finally, for any two genomes G_i and G_j, we compute their OG distance $D_{i,j}$ on basis of their OG pairs. Then we apply distance-based approaches of building trees, such as UPGMA, NJ and FM, to the matrix of OG distance between genomes for constructing genome trees of the input prokaryotic genomes.

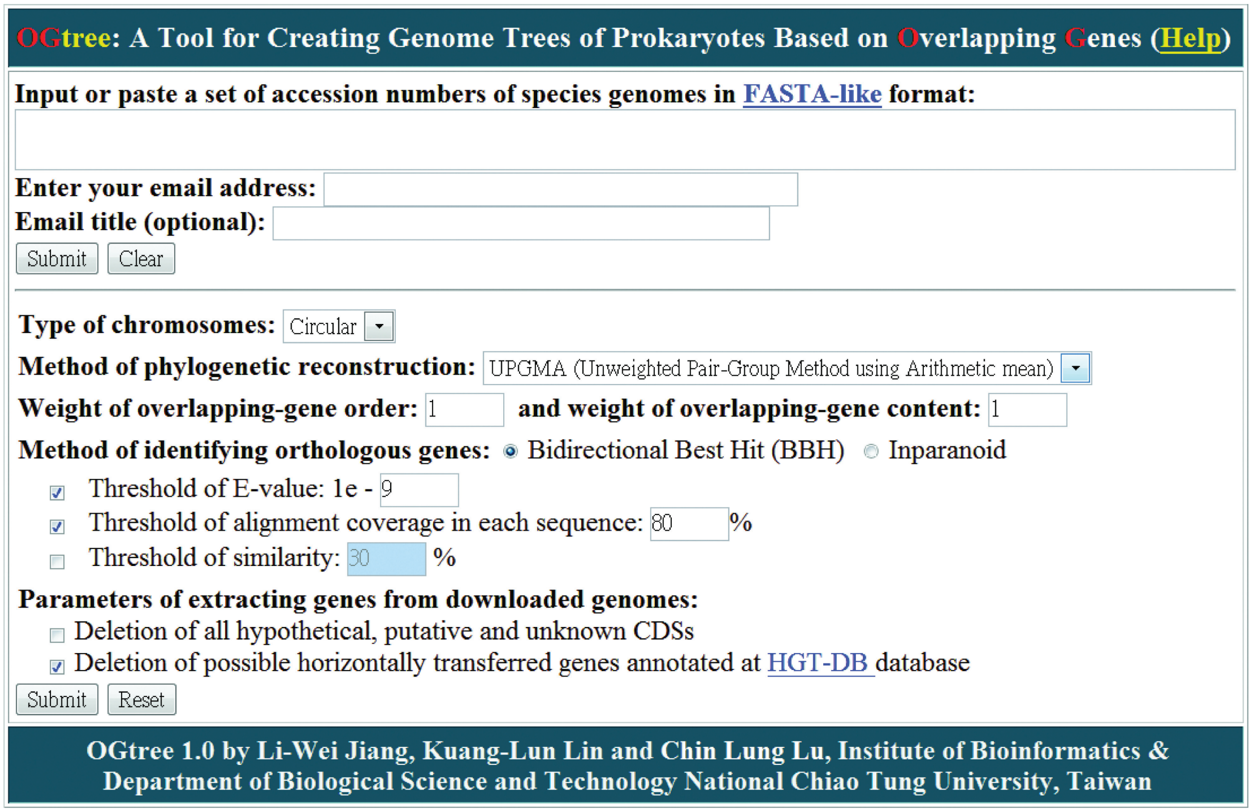
USAGE OF OGTREE

Input

OGtree provides an intuitive user interface as illustrated in Figure 1. The user can submit a job by entering or pasting a set of accession numbers of prokaryotic genomes in FASTA-like format (i.e. a single-line description beginning with a right angle bracket followed by a line of the accession number of a prokaryotic species), as well as an email address via which the user will be notified of the OGtree result when the submitted job is finished. In addition, the user is allowed to change or modify the default settings of all parameters, including the chromosomal type of input prokaryotic genomes (e.g. circular or linear), the distance-based method used by OGtree to build the genome tree (e.g. UPGMA, NJ or FM), the weights of OG content and order (e.g. any non-negative real numbers), the method used by OGtree to identify the orthologous genes (e.g. BBH or Inparanoid) and the thresholds of its *E*-value, alignment coverage in sequence and similarity, whether or not to delete all the CDSs whose translated products were annotated as hypothetical, putative and unknown proteins in the NCBI, and whether or note to delete all the CDS that were annotated as horizontally transferred genes at the HGT-DB database. For details, we refer the user to the help page of OGtree that can be easily reached via the link in the OGtree interface.

Output

In the output page, OGtree will show the OG distance matrix calculated according to the prokaryotic complete genomes downloaded from the NCBI (Figure 2), followed by a genome tree constructed using UPGMA, NJ or FM



OGtree: A Tool for Creating Genome Trees of Prokaryotes Based on Overlapping Genes (Help)

Input or paste a set of accession numbers of species genomes in FASTA-like format:

Enter your email address:

Email title (optional):

Type of chromosomes: Circular

Method of phylogenetic reconstruction: UPGMA (Unweighted Pair-Group Method using Arithmetic mean)

Weight of overlapping-gene order: 1 and weight of overlapping-gene content: 1

Method of identifying orthologous genes: ● Bidirectional Best Hit (BBH) ○ Inparanoid

☑ Threshold of E-value: 1e - 9

☑ Threshold of alignment coverage in each sequence: 80 %

□ Threshold of similarity: 30 %

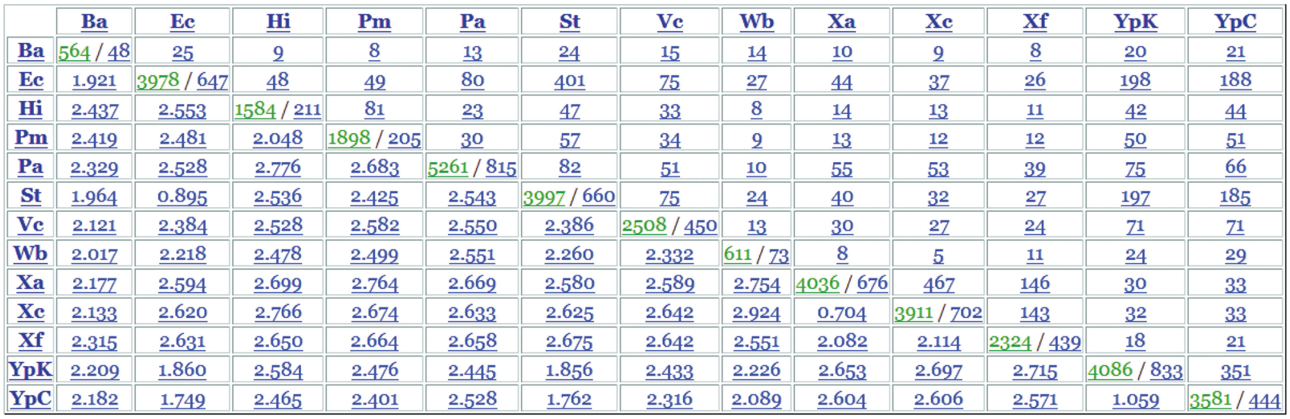
Parameters of extracting genes from downloaded genomes:

□ Deletion of all hypothetical, putative and unknown CDSs

☑ Deletion of possible horizontally transferred genes annotated at HGT-DB database

OGtree 1.0 by Li-Wei Jiang, Kuang-Lun Lin and Chin Lung Lu, Institute of Bioinformatics & Department of Biological Science and Technology National Chiao Tung University, Taiwan

Figure 1. OGtree web interface.



	Ba	Ec	Hi	Pm	Pa	St	Vc	Wb	Xa	Xc	Xf	YpK	YpC
Ba	564 / 48	25	9	8	13	24	15	14	10	9	8	20	21
Ec	1.921	3978 / 647	48	49	80	401	75	27	44	37	26	198	188
Hi	2.437	2.553	1584 / 211	81	23	47	33	8	14	13	11	42	44
Pm	2.419	2.481	2.048	1898 / 205	30	57	34	9	13	12	12	50	51
Pa	2.329	2.528	2.776	2.683	5261 / 815	82	51	10	55	53	39	75	66
St	1.964	0.895	2.536	2.425	2.543	3997 / 660	75	24	40	32	27	197	185
Vc	2.121	2.384	2.528	2.582	2.550	2.386	2508 / 450	13	30	27	24	71	71
Wb	2.017	2.218	2.478	2.499	2.551	2.260	2.332	611 / 73	8	5	11	24	29
Xa	2.177	2.594	2.699	2.764	2.669	2.580	2.589	2.754	4036 / 676	467	146	30	33
Xc	2.133	2.620	2.766	2.674	2.633	2.625	2.642	2.924	0.704	3911 / 702	143	32	33
Xf	2.315	2.631	2.650	2.664	2.658	2.675	2.642	2.551	2.082	2.114	2324 / 439	18	21
YpK	2.209	1.860	2.584	2.476	2.445	1.856	2.433	2.226	2.653	2.697	2.715	4086 / 833	351
YpC	2.182	1.749	2.465	2.401	2.528	1.762	2.316	2.089	2.604	2.606	2.571	1.059	3581 / 444

Figure 2. An example of OG distance matrix computed by OGtree for 13 γ-Proteobacteria: (lower triangle) OG distance between two genomes; (diagonal) numbers of genes (numerator) and OG pairs (denominator) identified in each genome; (upper triangle) number of orthologous OGs identified between two genomes.

method based on this OG distance matrix (Figure 4). OGtree also provides in the output page with a text file of computed OG distance matrix in the PYLIP format and a text file of constructed genome tree in the Newick format, so that the user can download them for post-processing analysis. In addition, OGtree provides each entry in the OG distance matrix with a link, through which the user can further review the details of related information, including genes and OG pairs extracted from each downloaded genome (in diagonal), orthologous OGs identified between each pair of input prokaryotic genomes (in upper-right triangle), and their orders in each of compared genomes (in lower-left triangle).

EXPERIMENTS

13 γ-Proteobacteria complete genomes

In this experiment, we selected 13 γ-Proteobacteria as the testing dataset that consists of *Buchnera aphidicola* (abbreviated as Ba, NC_002528), *Escherichia coli* (Ec, NC_000913), *Haemophilus influenzae* (Hi, NC_000907), *Pseudomonas aeruginosa* (Pa, NC_002516), *Pasteurella multocida* (Pm, NC_002663), *Salmonella typhimuriu* (St, NC_003197), *Vibrio cholerae* (Vc, NC_002505), *Wigglesworthia brevipalpis* (Wb, NC_004344), *Xanthomonas axonopodis* (Xa, NC_003919), *X. campestris* (Xc, NC_003902), *X. fastidiosa* (Xf, NC_002488), *Yersimia pestis* CO92 (YpC, NC_003143), and *Y. pestis* KIM (YpK, NC_004088). In addition, we used the phylogenetic trees constructed based on 16S rRNAs and concatenation of 205 orthologous proteins (17) as reference trees (Figure 3a and b) and compared the genome trees obtained by our OGtree to those phylogenetic tree (Figure 3c) predicted by Luo *et al.* (6). Basically, these two references have almost the same tree topology, just with a slight difference in the position of *V. cholerae*. The species of *V. cholerae* was placed as a neighbor of *P. aeruginosa* in the reference tree constructed using the concatenation of 205 proteins, whereas it was placed a little away from *P. aeruginosa* in the reference tree of 16S rRNAs.

As mentioned before, some misannotated genes may be included in the genomes of public databases. Therefore, we may exclude those CDSs annotated as being unknown, hypothetical or putative from each downloaded genome in our analysis, as done in ref. (6). However, we found that most of the CDSs in *W. brevipalpis* are currently annotated as unknown, hypothetical or putative, leading us to find no orthologous OG pair between *W. brevipalpis* and other species, if all these CDSs in *W. brevipalpis* are removed from our analysis. Here, instead of this method, we first removed those genes currently annotated as horizontally transferred genes at the HGT-DB database (14) and then applied more stringent criteria of identifying putative orthologous genes by using BBH and setting the parameters with at least 80% of each authentic CDS sequence involved in the alignment and a minimum *E*-value of 10^{-9}.

Consequently, the NJ and FM trees (Figures 4b and c, respectively) we obtained using OGtree have the same tree topology, which slightly differ from the one in the UPGMA tree (Figure 4a) with respect to the positions of *W. brevipalpis* and *B. aphidicola*. The two endosymbionts of *W. brevipalpis* and *B. aphidicola* were placed as neighbor taxa in the NJ and FM trees, whereas they were as a sister group in the UPGMA tree.

In the comparison of the phylogenetic trees inferred by Luo *et al.* (6), our genome trees show more precise and robust phylogenies for the completely sequenced genomes of 13 γ-Proteobacteria. For instance, the topology of the UPGMA tree (Figure 4a) we constructed here based on the OG distance is completely consistent with that in the reference tree based on 16S rRNAs (Figure 3a), and nearly consistent with that in the reference tree constructed using the concatenation of 205 proteins (Figure 3b). It is worth mentioning that the two endosymbionts *W. brevipalpis* and *B. aphidicola* were separated from each other in the UPGMA tree (Figure 3c) constructed by Luo *et al.* (6). In contrast, *W. brevipalpis* and *B. aphidicola* in our UPGMA tree, as well as in both reference trees, were

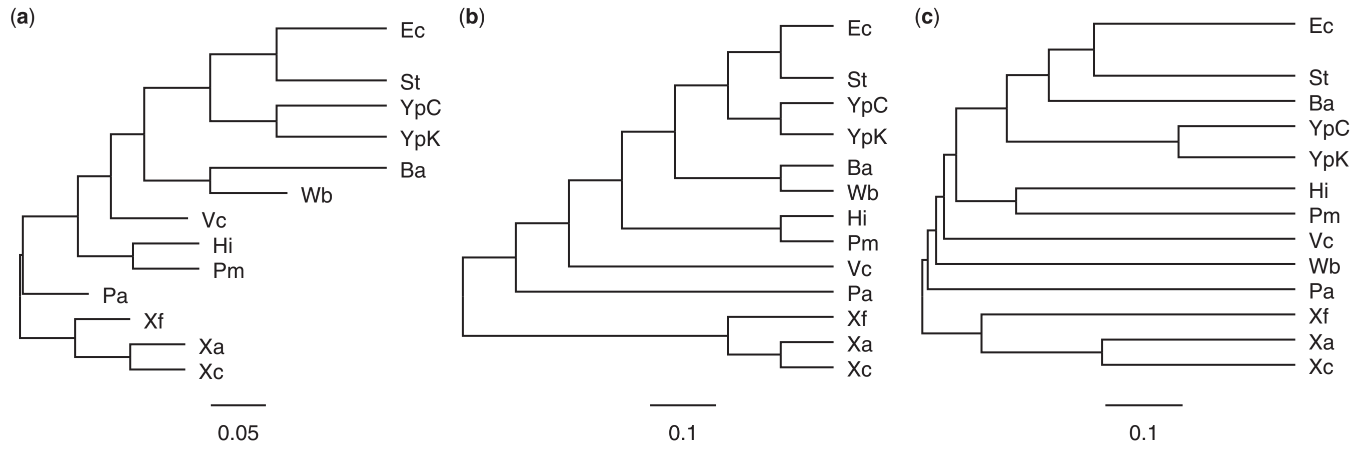


Figure 3. (**a**) NJ tree constructed using 16S rRNA sequences, (**b**) NJ tree constructed based on concatenation of 205 proteins and (**c**) UPGMA tree constructed by Luo *et al.* (6) for 13 γ-Proteobacteria, where the reference trees shown in (a) and (b) were adapted from refs (6) and (17), respectively.

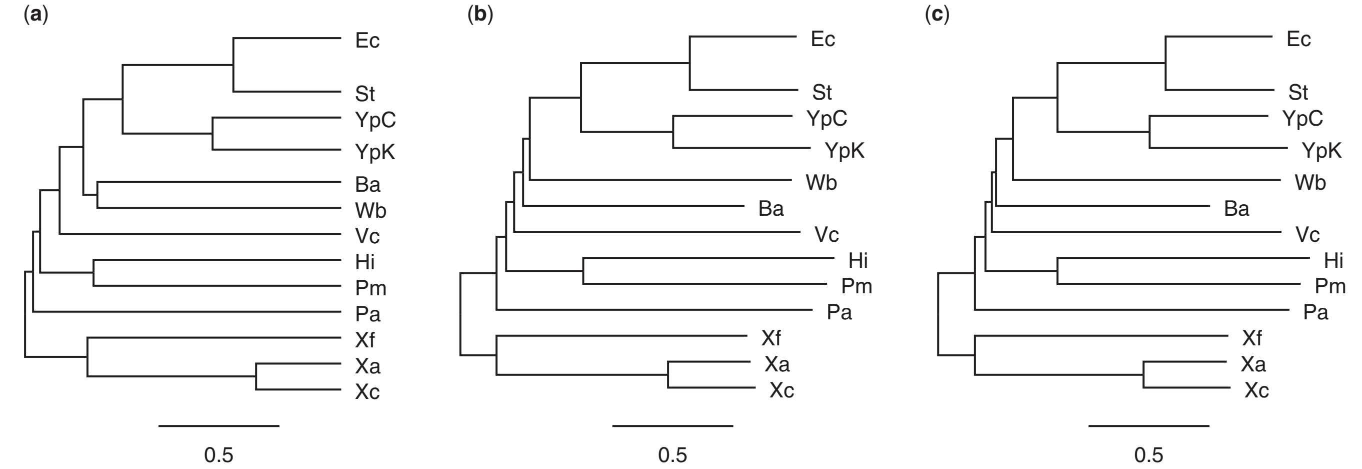


Figure 4. (**a**) UPGMA tree, (**b**) NJ tree and (**c**) FM tree constructed by OGtree for 13 γ-Proteobacteria.

placed as a sister group, suggesting that there should be a common origin for these two species both of which are symbiotic and have reduced genomes.

Among the three tree-building methods in our experiment, the UPGMA method produced a much more congruent genome tree compared to both the NJ and FM methods, if they were based on the OG distance we defined in this study. This characteristic was also pointed out by Luo *et al.* (6,7) in their studies only on the basis of the content of OG pairs. It has been reported that evolution of OGs occurs at a universal mutation rate across bacterial genomes (8,9). Perhaps due to this property, the UPGMA method is more suitable for the reconstruction of phylogenies particularly based on OG pairs, when compared to the NJ and FM methods.

18 Proteobacteria complete genomes

In the second experiment, we reconducted the above experiment but with including additional two α-Proteobacteria, *Caulobacter crescentus* (abbreviated as Cc, NC_002696) and *Rickettsia conorii* (Rc, NC_003103), and three β-Proteobacteria, *Nitrosomonas europaea* (Ne, NC_004757), *Neissenia meningitidis* MC58 (NmM, NC_003112) and *N. meningitidis* Z2491 (NmZ, NC_003116). In the UPGMA tree constructed by Luo *et al.* (7), as was shown in Figure 5a, the species *N. europaea*, a β-Proteobacteria, was separated from the other two β-Proteobacteria *N. meningitidis* MC58 and *N. meningitidis* Z2491 and was placed in the group containing all 13 γ-Proteobacteria. In contrast, all these three β-Proteobacteria in our UPGMA tree was placed as a sister group, as illustrated in Figure 5b. Particularly, the testing α-, β- and γ-Proteobacteria correctly form three monophyletic clades in our UPGMA tree.

SUMMARY

OGtree is a web-based tool for reconstructing genome trees of prokaryotes according to their pairwise OG distance. The OG distance defined here is based on a combination of OG content and orthologous OG order. In this study, we have demonstrated that our OGtree was

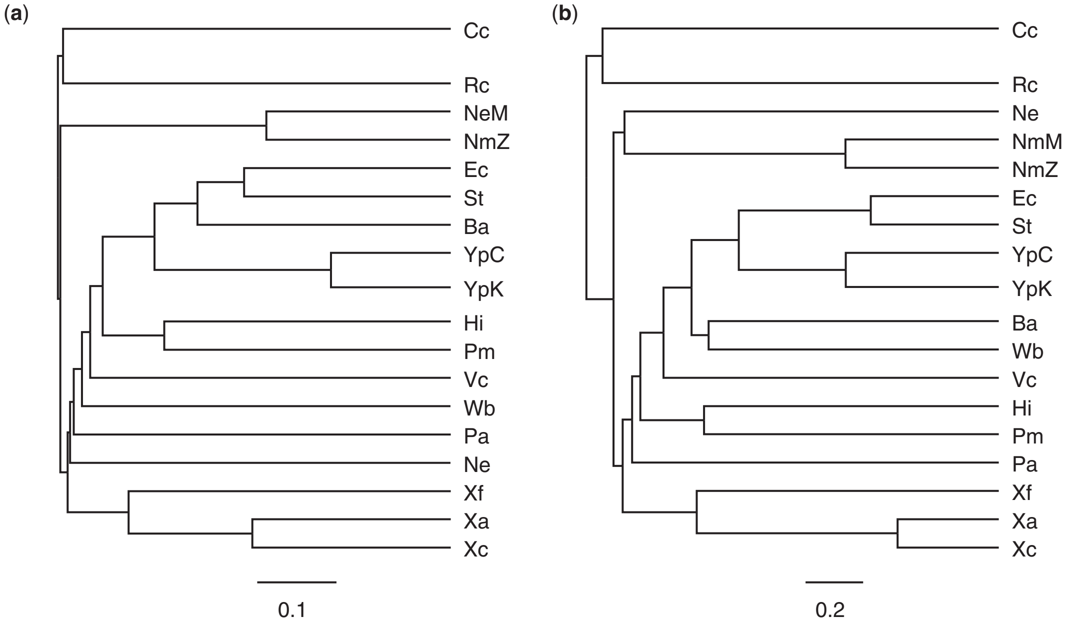


Figure 5. (**a**) UPGMA tree inferred by Luo *et al.* (7) and (**b**) UPGMA tree produced by our OGtree using 18 Proteobacteria genomes.

able to construct more precise and robust genome trees for some prokaryotic genomes. Therefore, we believe that our OGtree can provide interesting insights into the study of evolutionary relationships of completely sequenced prokaryotic genomes.

ACKNOWLEDGEMENTS

This work was supported in part by National Science Council of Republic of China under grant NSC96-2221-E-009-126 and in part by the ATU plan of MOE. Funding to pay the Open Access publication charges for this article was provided by National Science Council of Republic of China under grant NSC96-2221-E-009-126.

Conflict of interest statement. None declared.

REFERENCES

1. Snel,B., Bork,P. and Huynen,M.A. (1999) Genome phylogeny based on gene content. *Nat. Genet.,* **21**, 108–110.
2. Snel,B., Huynen,M.A. and Dutilh,B.E. (2005) Genome trees and the nature of genome evolution. *Ann. Rev. Microbiol.,* **59**, 191–209.
3. Blanchette,M., Kunisawa,T. and Sankoff,D. (1999) Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.,* **49**, 193–203.
4. Sankoff,D. (1999) Genome rearrangement with gene families. *Bioinformatics,* **15**, 909–917.
5. Belda,E., Moya,A. and Silva,F.J. (2005) Genome rearrangement distances and gene order phylogeny in γ-Proteobacteria. *Mol. Biol. Evol.,* **22**, 1456–1467.
6. Luo,Y., Fu,C., Zhang,D.Y. and Lin,K. (2006) Overlapping genes as rare genomic markers: the phylogeny of γ-Proteobacteria as a case study. *Trends Genet.,* **22**, 593–596.
7. Luo,Y., Fu,C., Zhang,D.Y. and Lin,K. (2007) BPhyOG: an interactive server for genome-wide inference of bacterial phylogenies based on overlapping genes. *BMC Bioinform.,* **8**, 266.
8. Fukuda,Y., Nakayama,Y. and Tomita,M. (2003) On dynamics of overlapping genes in bacterial genomes. *Gene,* **323**, 181–187.
9. Johnson,Z.I. and Chisholm,S.W. (2004) Properties of overlapping genes are conserved across microbial genomes. *Genome Res.,* **14**, 2268–2272.
10. Fukuda,Y., Washio,T. and Tomita,M. (1999) Comparative study of overlapping genes in the genomes of *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. *Nucleic Acids Res.,* **27**, 1847–1853.
11. Krakauer,D.C. (2000) Stability and evolution of overlapping genes. *Evol. Int. J. Organ. Evol.,* **54**, 731–739.
12. Sakharkar,K.R., Sakharkar,M.K., Verma,C. and Chow,V.T. (2005) Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*. *Int. J. Syst. Evol. Microbiol.,* **55**, 1205–1209.
13. Rogozin,I.B., Spiridonov,A.N., Sorokin,A.V., Wolf,Y.I., Jordan,I.K., Tatusov,R.L. and Koonin,E.V. (2002) Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.,* **18**, 228–232.
14. Garcia-Vallve,S., Guzman,E., Montero,M.A. and Romeu,A. (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.,* **31**, 187–189.
15. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science,* **278**, 631–637.
16. Hulsen,T., Huynen,M.A., deVlieg,J. and Groenen,P.M. (2006) Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.,* **7**, 4.
17. Lerat,E., Daubin,V. and Moran,N.A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the γ-Proteobacteria. *PLoS Biol.,* **1**, E19.